Current Synthetic and Systems Biology
ISSN: 2332-0737

Runthala, Curr Synthetic Sys Biol 2015, 3:2
DOI: 10.4172/2332-0737.1000121

Research Article Open Access

Non-Linear and Misleading Template Scoring Criteria: Root Cause of Protein Modelling Inaccuracies

Ashish Runthala*
Department of Biological Sciences,Birla, Institute of Technology & Science, Pilani, India

Abstract

Template based protein modelling is currently the most accurate as well as trustworthy method for predicting the correct protein conformations to bridge the constantly increasing gap between the number of experimentally solved protein structures and the count of protein sequences. Our best knowledge based prediction algorithms employing the templates are not highly proficient of consistently selecting the best scoring template(s) to construct a highly accurate protein model. Mutually contrary nature of generic and currently employed template assessment and selection scores further makes this essential modelling step a very tricky and fluky business. Precisely, the article briefly investigates and justifies the impact of fundamentally allowed degree of freedom of a template selection measure on the accuracy of constructed protein models. Several logical guidelines, normally overlooked in a protein modelling task, are analyzed and should be routinely considered. A more reliable and robust scoring measure is thus mandatorily required to select the best possible available template for constructing the most accurate target conformation.

Keywords: Protein; CASP; Modelling; Template; Assessment

Introduction

Functional study of proteins is based on the highly accurate knowledge of their structural details. Structure determination methodologies, aimed at constructing accurate conformations, face several technical and monetary limitations. Protein modelling algorithms hereby come for the rescue to quickly predict highly accurate structures [1]. Modelling accuracy of a protein sequence depends on the degree of near-native proximity of a predicted model [2]. A highly accurate Template Based Modelling (TBM) algorithm [3] employs the structural information of solved protein structures (*templates*), available in the Protein Data Bank (PDB), to maximally span the target (*Considered protein sequence for modelling*) [4]. Gapped or unaligned segments in such target-template alignments are possibly the results of insertions, deletions (INDELs) primarily caused due to evolutionary pressure and are modelled through a couple of means. Such segments are normally modelled through the physical principle of protein folding for building the lowest energy confirmation. Distantly related or dissimilar templates are also employed for modelling the target chunk, not spanned by the selected templates, to construct an overall model [5]. Algorithms employing the correct as well as biologically significant templates have been proven to construct fairly accurate models. Major steps of a generic TBM algorithm include several steps, amongst which template identification and selection step is of paramount importance. The selected template(s) are then aligned with the target sequence to construct a model [6] and such predictions are normally employed for several cellular applications [7]. tools also employ PDB culling [8-19] at a sequence identity threshold and that is actually the mutual comparison of templates, which may yield a template as high-scoring hit, although it is structurally too distant from the structural topology of the target sequence.

To solve most of the modelling errors caused due to an incorrect structure, not functionally or biologically related to the target, a reliable set of template(s) is thus normally selected through the following scoring measures. However, they are usually antagonistic and they do not unanimously select the best template as the top ranked hit consistently. The degree of multi-dimensional scoring schemes forces us to follow the best possible scoring scheme, which is mostly the consideration of the E-value scores. Therefore, false positive and spurious templates with significant homoplasic sequence similarity to target sequence are not eminently, reliably distinguished and filtered out from the correct set of actually relevant templates. This concept is well illustrated in Tables 1 and 2 which enlist the modelling accuracy of randomly chosen CASP8 targets T0423 and T0428 through several significant templates searched by HHPred. Targets T0423 and T0428 encode a sequence length of 110 and 267 residues respectively. The near-native accuracy of these target models was assessed, as per the structural domain information employed by CASP, respectively through 97 (*2-98*) and 229 (*20-248*) residue lengths.

Target-template length difference

An ideal template is expected to encode the same number of residues as the target sequence, as expected. It works fairly well against a single domain template employed for a short length target sequence. However, it is usually quite hard to see such case due to domain insertion, duplication or deletion in the templates and so a stretch of a structural domain of a template may sometimes provide the best structural information for a target sequence. Reliability of a hit to be the actually best template for a target sequence thus becomes a dilemma.

Substitution matrices

Substitution matrices, scoring and justifying feasibility of the aligned residue substitutions in a target-template alignment, are considered credible. However, it mostly becomes an erroneous case when a template residue say Glycine (G) come up in mutation over the earlier residue Alanine (A) and on alignment with the target residue G at that locus, it would become the false positive sequence identity. Such

***Corresponding author:** Ashish Runthala, Department of Biological Sciences,Birla Institute of Technology & Science, Pilani, India, Tel: +91-7597971146; Fax: +91-1596-244183; E-mail: ashish.runthala@gmail.com

Received April 13, 2015; **Accepted** May 09, 2015; **Published** May 11, 2015

Citation: Runthala A (2015) Non-Linear and Misleading Template Scoring Criteria: Root Cause of Protein Modelling Inaccuracies. Curr Synthetic Sys Biol 3: 121. doi:10.4172/2332-0737.1000121

Template	Resolution	Length	Sequence Identity	Average gap length	BLOSUM score	Mismatch residues	Coverage span	TM_Score	GDT-TS	RMSD
2OTM	1.85	152	33.77	1.3	-799.09	64.94	98.72	0.937	90.068	1.077
1PF5	2.50	130	26.57	9.79	-873.21	63.64	99.36	0.572	45.377	2.066
2B33	2.30	127	27.56	10.25	-877.46	62.19	100	0.514	36.815	2.63
2CVL	1.65	124	30.71	11.43	-886.62	57.86	99.36	0.561	45.377	1.939
1QD9	1.70	124	29.29	12.14	-894.48	58.57	99.36	0.541	44.178	1.768
1QU9	1.20	126	29.79	10.64	-881.32	59.57	99.36	0.526	41.781	2.302
2CWJ	3.60	116	26.47	14.71	-914.58	58.82	98.72	0.52	38.527	2.368
2DYY	2.60	125	28.47	11.03	-883.74	60.5	99.36	0.506	36.986	2.429
1XRG	2.20	125	27.76	11.03	-884.47	61.21	99.36	0.496	34.76	2.697
2IG8	1.90	142	26.85	9.4	-869.94	63.76	96.32	0.458	32.877	2.837
1X25	2.00	126	30.5	10.64	-880.91	58.87	98.72	0.447	32.363	2.235
2EWC	2.15	120	17.39	13.77	-906.86	68.84	100	0.109	5.479	2.881

Table 1: Inconsistent template selection scoring results, showing their non-linear relationship with the most credible GDT-TS and TM_Score measure, for the CASP8 target T0423 encoding 110 residues.

homoplasy, along with discrete behavior of context specific localization of residues at specific sites in a protein, makes the reliability of a selected template completely questionable. Therefore, such algorithms have been improved several times to consider the structural information along with the sequence context information [20-23]. It is well understood that sequence similarity is a good score to select reliable templates, however it is not always correct and the most similar hit is not always the best one. Still, it has been consistently employed in most of the modelling algorithms and we are far from utilizing it as the best scoring measure for a target sequence.

Residue composition

Differences in the proportions of encoded amino acids between target and a template sequence form the basis of this scoring scheme. Lower is the difference or more closer is the observed residue composition proportions in the target and template sequences irrespective of their alignment, higher is the affirmed reliability of the template being phylogenetically closer and credible for the target. CASP8 algorithms including CADCMLAB [24] and COMA [25], CASP9 algorithms including DCLAB [24] and FLYPRED [26] and CASP10 algorithms including samcha-server [27] and TSAILAB [28] employed such a scoring measure to rank the templates against the considered target sequence [29].

Sequence identity

This measure, though being considered as a highly reliable scoring criterion, is also impaired by homoplasy. Structural and functional similarity of templates is thus normally used to select the best template amongst the set of redundant hits for a target sequence. This structural similarity is also normally employed for phylogenetic study of the protein structures [30]. However, when several hits share an almost equal sequence identity score with the target, their credibility seems to be doubtful and it may become difficult to select the most similar one. It was well realized by Jones-UCL in CASP9, and here the template culling step, to keep the one most significant template amongst similar structures for a target, further complicates the process. Such a template culling step further poses a new challenge. It is because template selection step is employed to screen them against the target and their mutual comparison may exclude the actually closest and reliable hit. This scoring scheme was used by several groups including CaspIta [31] and COMA [25] in CASP8, ATOME2_CBS [32] and FAMSD [33] in CASP9, and ATOME2_CBS [32] and CASPita [31] in CASP10.

Coverages

Coverage span, seemingly a reliable measure, is dependent on alignment constraints and the employed scoring scheme. A shorter and a longer sequence even if aligned together, for an almost complete coverage span, make the corresponding template selection lucrative. However, comparative analysis of the target against another template with same coverage span and a higher sequence identity makes the later hit a favorable choice. However in another case, if an actually correct template is evolutionarily closer to the target and has lesser coverage span, its selection again becomes a questionable dilemma for the person. This measure has been used by several algorithms including BAKER-GINZU and PRO-SP3-TASSER [34] in CASP8, Firestar [35] and GSmetadisorder [36] in CASP9 and CASPita [31] and HHPred [10] in CASP10.

Alignment score, e-value

These scores depend on the quality of alignment. An Alignment score and an E-value, the chance to expect the same template in the PDB database, are good scoring factors to correctly discriminate between highly correct and worse templates. Through the target-template HMM profile comparison, the alignment score of a template hit is also computed. However, these measures also fail to precisely select the best template from the pretty similar set of almost equally scoring hits and so there is still a need to develop other significant assessment measures for a target to model a highly accurate protein conformation [37]. In CASP8, FAMSD, FAMSSEC, sbtJ, Yuan-Chen-Kihara and ZHOU-SPARKS-X majorly relied on this scoring scheme. This scoring scheme has been used by many algorithms including Distill [38] in CASP8, Distill [38] and TASSER [39] in CASP9 and TASSER [39] and Distill [38] in CASP10.

Resolution

For a protein, it scores the experimental quality of data obtained from the crystal. On the basis of structurally similar topology of proteins, a crystal results in a diffraction pattern and a perfect highly ordered structure shows a resolution score of 1Å. It is normally believed that high resolution structures solved by X-Ray experimental methodology are the perfect ones. However, it is not always applicable for selecting the best set of templates. A template protein, even with a lower resolution, may still provide the structural information for several target residues, not spanned by the already selected templates, and may thus be fairly reliable conformation.

Template	Resolution	Length	Sequence Identity	Average gap length	BLOSUM score	Mismatch residues	Coverage Span	TM_Score	GDT-TS	RMSD
1XQ9	2.58	241	60.24	5.12	-829.99	34.65	96.63	0.977	94.932	0.886
1E59	1.30	239	53.75	5.53	-834.43	40.71	100	0.962	91.376	1.225
1YJX	2.80	245	48.83	5.47	-833.45	45.7	97.41	0.951	88.537	1.397
1YFK	2.70	243	48.63	5.88	-836.98	45.49	97.04	0.951	87.882	1.343
1T8P	2.50	249	45.74	5.81	-836.67	48.45	98.17	0.934	84.607	1.549
2H4Z	2.00	255	46.74	4.98	-829.54	48.28	99.64	0.909	81.223	1.711
2A9J	2.00	253	46.54	5	-829.6	48.46	100	0.909	81.004	1.676
2H4X	1.85	255	46.74	4.98	-829.54	48.28	99.64	0.908	81.223	1.722
2H52	2.00	255	46.74	4.98	-829.54	48.28	99.64	0.907	81.004	1.663
2F90	2.00	254	46.45	5.18	-831.24	48.37	99.64	0.907	80.667	1.699
1FZT	NA	211	44.77	11.72	-888.03	43.51	98.13	0.809	69.105	1.687
2P30	1.85	177	30.18	20.27	-961.17	49.55	98.88	0.603	42.467	2.629
2OWD	1.65	171	30.14	21.92	-975.3	47.95	97.75	0.602	42.576	2.69
2P78	1.75	171	29.22	21.92	-975.36	48.86	97.75	0.6	43.122	2.483
2P75	1.70	171	30.14	21.92	-975.31	47.95	97.75	0.6	41.476	2.675
1V37	1.40	171	29.22	21.92	-975.3	48.86	97.75	0.599	37.079	2.346
2PA0	2.30	171	28.31	21.92	-975.43	49.77	98.13	0.598	42.476	2.618
2EKB	1.70	171	30.14	21.92	-975.29	47.95	97.75	0.597	41.157	2.7
2P2Z	1.75	171	29.22	21.92	-975.38	48.86	97.75	0.594	42.14	2.48
2P6O	1.65	171	29.68	21.92	-975.32	48.4	97.75	0.592	42.467	2.596
2P9Y	1.85	171	29.22	21.92	-975.37	48.86	97.75	0.591	42.467	2.462
2P6M	1.90	171	29.22	21.92	-975.38	48.86	97.75	0.591	41.266	2.449

Table 2: Inconsistent template selection scoring results, showing their non-linear relationship with the most credible GDT-TS and TM_Score measure, for the CASP8 target T0428 encoding 267 residues.

These measures, enlisted in the Tables 1 and 2 ordered as per TM_Score [37] accuracy, are quite heterogeneous and their reliability varies a lot. Errors due to the selection of seemingly reliable but actually evolutionarily distant hit should thus be tackled properly, as recently tried by several modelling algorithms [40]. The servers trying to solve this issue through computation of multiple sequence profiles are highly laborious, time-consuming and still cannot predict highly accurate models consistently [41-43].

Suggested Strategy

The template search and selection step of a routinely employed protein modelling algorithm should be properly screened. Several knowledge based expert guidelines, as enlisted and explained in logistically correct order below, should thus be routinely considered to select the best template(s) for a target sequence.

Maximum informative MSA profile

Iterative template search rounds are often employed by several algorithms including HHPred [10] to search the significantly relevant set of templates for a target sequence. Such a maximum allowed iteration parameter, although computationally expensive, fairly correlates and considers even the distantly related hits for a target sequence and should thus be normally employed. It reasonably evaluates the evolutionarily consensus probability of residue substitutions across the target sequence in the screened list of hits to prioritize and accurately rank the scoring of correctly related templates.

E-value threshold

A hit with a considerably low E-value score is normally considered as a good template for a target sequence. This concept quite reasonably selects the best hit with the lowest E-value score for a target sequence. However, the same relationship

can never be extrapolated to other meaningful templates as a hit with a very bad E-value score might still be sequentially divergent as well as structurally and functionally relevant one for a target sequence. Hence, solely discarding templates for their lower E-value scores is not normally advised.

Score secondary structure of target

A target sequence might share too much sequence divergence with the selected functionally similar templates and still be excellent structural resource. Hence as per the constructed reasonably correct alignment, similarity of predicted secondary structure of target sequence chunks and the template segments provides reliable homology information. This constraint also assists the construction of a reasonably accurate target-template alignment.

Local as well as global alignment consideration

Protein sequence information is significantly lost through inaccurate localization of gaps especially while constructing an optimally scoring and biologically meaningful alignment. Hence, gaps must be carefully crosschecked in the target-template alignment. Similarly, longer gap segments more than 5 gaps should be avoided, especially if they not at the periphery, as *ab-initio* modelling of such chunks might disturb the orientation and topology of adjoining residues especially if they encode a secondary structure element. Therefore, an alignment optimally placing the residues, both in terms of their local and global functional significance, should be employed for a logistically correct modelling of the target sequence. The best possible biologically meaningful alignment might not always be the one with mathematically best score and rather it could be a sub-optimal one with a comparatively inferior score.

Functional significance of templates

A target sequence normally encodes atleast one functional domain, although it might be sequentially and structurally continuous or

discontinuous. Hence, a target sequence must be screened for the plausible availability and localization of such domains through several databases including PFAM and CDD and then the functionally similar Homologues and Orthologoues must be considered as reliable hits through other scoring measures. Such a consideration of structurally and functionally significant sequence information normally involves the exquisite evolutionarily reliable sequence information of templates the best possible way and thus assists us to predict highly accurate near-native protein models for both the conserved local structural segments and the complete model altogether.

Employing all culled PDBs

Consider all the culled PDB structures along with the selected functionally similar and reliable representative hits for selecting the best possible set of templates for a target sequence. The culled PDB might actually be evolutionarily and structurally closest to the target sequence and hence the complete set of related PDB structures must be considered for selecting the best set of templates.

Fixing the best set of templates

Best possible set of mutually and structurally complementary templates is essential to model a highly accurate protein structure. The discussed scoring measures and template selection or consideration constraints must thus be carefully employed, through correctly computed pairwise and multiple sequence alignments, to fix up the best possible templates for maximally spanning the target sequence. The best hit, scored significantly with majority of the aforementioned measures, must be thus employed to seed the construction of a highly accurate MSA. This MSA should then be employed for screening the hits to maximally span the target.

Conclusion

The template search and selection criterion, being the major armature to ultimately build the highly reliable models, needs a well developed template ranking system. Selecting the reliable templates is thus the supreme prerequisite to construct highly accurate protein models. CASP Server models are therefore usually pretty poor topology predictions and are not highly accurate compared to the well justified human models. A robust template selection algorithm, encompassing the best of these scoring measures, is thus required to significantly distinguish the actually relevant templates from the spurious hits and thus solve modelling errors caused due to consideration of incorrect template(s).

Discussion

Most of the template search and selection criteria seem to be parallel or mutually convergent with consideration of their benchmarked weights. A robust algorithm with optimally weighed consideration of most of these measures is thus required to reliably rank the credibility of a template. However, it is obvious that any such template ranking algorithm will fail completely when an assigned weight results in a false positive ranking of templates. Weighting increases the credibility of a selection measure much more than others and the marginal change of the weighted factor significantly suppresses the noteworthy weights of other template scoring measures. Template scoring results of the significant hits searched by HHPred [10] for the CASP8 targets T0423 and T0428, enlisted in the Tables 1 and 2, clearly prove this discussed problem. All the aforementioned template selection measures are enlisted in these tables and their scores are not always parallel to the TM_Score [37], computed against the actual native conformation. A template solely selected on the basis of a single scoring measure may not be the best structural hit always and so a more reliable template scoring measure, statistically too robust, is mandatorily required to definitely pave our way for developing a consistently successful modelling algorithm.

References

1. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. (2000) The protein data bank. Nucleic Acid Research 28: 235-242.
2. Almerico AM, Tutone M, Pantano L, Lauria A (2013) A3 adenosine receptor: Homology modeling and 3D-QSAR studies. Journal of Molecular Graphics and Modelling 42: 60-72.
3. Rost B, Liua J, Naira R, Wrzeszczynski KO, Ofran Y (2003) Automatic prediction of protein function. Cellular and Molecular Life Sciences 60: 2637–2650.
4. Zhang Y, Skolnick J (2004) The protein structure prediction problem could be solved using the current PDB library. The Proceedings of the National Academy of Sciences USA 102: 1029–1034.
5. Bray JE, Todd AE, Pearl FM, Thornton JM, Orengo CA (2000) The CATH Dictionary of Homologous Superfamilies (DHS): a consensus approach for identifying distant structural homologues. Protein Engineering Design & Selection 13: 153-165.
6. Rychlewski L, Jaroszewski L, Li W, Godzik A (2000) Comparison of sequence profiles. Strategies for structural predictions using sequence information. Protein Science 9: 232-241.
7. Hanson B, Westin C, Rosa M, Grier A, Osipovitch M, et al. (2014) Estimation of protein function using template-based alignment of enzyme active sites. BMC Bioinformatics 15: 87.
8. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang W, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25: 3389-3402
9. Gonzalez MW, Pearson WR (2010) Homologous overextension: A challenge for iterative similarity searches. Nucleic Acids Research 38: 2177-2189.
10. Soding J (2005) Protein homology detection by HMM-HMM comparison. Bioinformatics 21: 951-960.
11. Hildebrand A, Remmert M, Biegert A, Soding J (2009) Fast and accurate automatic structure prediction with HHpred. Proteins: Structure, Function and Bioinformatics 77: 128-132.
12. Sadowski MI, Jones DT (2007) Benchmarking template selection and model quality assessment for high-resolution comparative modeling. Proteins: Structure, Function and Bioinformatics 69: 476-485.
13. Angermüller C, Biegert A, Söding J (2012) Discriminative modelling of context-specific amino acid substitution probabilities. Bioinformatics 28: 3240-3247.
14. Ma J, Wang S (2014) Algorithms, applications, and challenges of protein structure alignment. Advances in Protein Chemistry and Structural Biology 94: 121-175.
15. Kanou K, Iwadate M, Hirata T, Terashi G, Umeyama H, et al. (2009) FAMSD: A powerful protein modeling platform that combines alignment methods, homology modeling, 3D structure quality estimation and molecular dynamics. Chemical & Pharmaceutical Bulletin 57: 1335-1342.
16. Lobley A, Sadowski MI, Jones DT (2009) pGenTHREADER and pDomTHREADER: new methods for improved protein fold recognition and superfamily discrimination. Bioinformatics 25: 1761-1767.
17. Xu D, Zhang Y (2012) Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. Proteins: Structure, Function and Bioinformatics 80:1715-1735.
18. Gniewek P, Kolinski A, Kloczkowski A, Gront D (2014) BioShell-Threading: versatile Monte Carlo package for protein 3D threading. BMC Bioinformatics 15: 22.
19. Wang G, Dunbrack RL Jr (2005) PISCES: recent improvements to a PDB sequence culling server. Nucleic Acids Research 33: 94-98.
20. Overington J, Donnelly D, Johnson MS, Sali A, Blundell TB (1992) Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. Protein Science 1: 216-226.

21. Lobanov MY, Finkelstein AV (2010) Analogy-based protein structure prediction: III. Optimizing the combination of the substitution matrix and pseudopotentials used to align protein sequences with spatial structures. Molecular Biology 44: 109-118.

22. Shi J, Blundell TL, Mizuguchi K (2001) FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. Journal of Molecular Biology 310: 243-257.

23. Teodorescu O, Galor T, Pillardy J, Elber R (2004) Enriching the sequence substitution matrix by structural information. Proteins: Structure, Function and Bioinformatics 54: 41-48.

24. Del Carpio CA, Carbajal JC (2002) Folding Pattern Recognition in Proteins Using Spectral Analysis Methods. Genome Informatics 13: 163-172.

25. Venclovas C, Margelevicius M (2009) The use of automatic tools and human expertise in template-based modeling of CASP8 target proteins. Proteins: Structure, Function and Bioinformatics 77: 81-88.

26. Björkholm P, Daniluk P, Kryshtafovych A, Fidelis K, Andersson R, et al. (2009) Using multi-data hidden Markov models trained on local neighborhoods of protein structure to predict residue-residue contacts. Bioinformatics 25: 1264-1270.

27. Jeong CS, Kim D (2012) Reliable and robust detection of coevolving protein residues. Protein Engineering Design and Selection 25: 705-713.

28. Joo H, Chavan AG, Phan J, Day R, Tsai J (2012) An amino acid packing code for α-helical structure and protein design. Journal of Molecular Biology 419: 234-254.

29. Rykunov D, Fiser A (2007) Effects of amino acid composition, finite size of proteins, and sparse statistics on distance-dependent statistical pair potentials. Proteins: Structure, Function, and Bioinformatics 67: 559-568.

30. Sabharwal NS, Runthala A (2014) Functional protein domains evolve very specifically over mutations. Journal of Proteomics and Genomics. 1: 102.

31. Sirocco F, Tosatto SC (2008) TESE: Generating specific protein structure test set ensembles. Bioinformatics 24: 2632-2633.

32. Pons JL, Labesse G (2009) @TOME-2: a new pipeline for comparative modeling of protein-ligand complexes. Nucleic Acids Research 37: 485-491.

33. Kanou K, Iwadate M, Hirata T, Terashi G, Umeyama H, et al. (2009) FAMSD: A powerful protein modeling platform that combines alignment methods, homology modeling, 3D structure quality estimation and molecular dynamics. Chemical & Pharmaceutical Bulletin (Tokyo) 57: 1335-1342.

34. Zhou H, Pandit SB, Skolnick J (2009) Performance of the Pro-sp3-TASSER server in CASP8. Proteins: Structure, Function, and Bioinformatics 77: 123-127.

35. Lopez G, Maietta P, Rodriguez JM, Valencia A, Tress ML (2011) firestar—advances in the prediction of functionally important residues. Nucleic Acids Research 39: 235-241.

36. Deng X, Eickholt J, Cheng J (2012) A comprehensive overview of computational protein disorder prediction methods. Molecular Biosystems 8: 114-121.

37. Xu J, Zhang Y (2010) How significant is a protein structure similarity with TM-score=0.5? Bioinformatics 26: 889-895.

38. Baú D, Martin AJM, Mooney C, Vullo A, Walsh I, et al. (2006) Distill: a suite of web servers for the prediction of one-, two- and three-dimensional structural features of proteins. BMC Bioinformatics 7: 402.

39. Zhang Y, Arakaki AK, Skolnick J (2005) TASSER: an automated method for the prediction of protein tertiary structures in CASP6. Proteins: Structure, Function, and Bioinformatics 61: 91-98.

40. Thompson J, Baker D (2011) Incorporation of evolutionary information into Rosetta comparative modeling. Proteins: Structure, Function, and Bioinformatics 79: 2380-2388.

41. Margelevicius M, Venclovas C (2010) Detection of distant evolutionary relationships between protein families using theory of sequence profile-profile comparisons. BMC Bioinformatics 11: 89.

42. Runthala A, Chowdhury S (2013) Protein Structure Prediction: Are we there yet?. Knowledge Based Systems in Biomedicine and Computational Life Science, Studies in Computational Intelligence, 450: 79-115.

43. Runthala A (2012) Protein Structure Prediction: Challenging targets for CASP10. Journal of Biomolecular Structure and Dynamics 30: 607-615.